NAVAL HEALTH RESEARCH CENTER

RUNNING PERFORMANCE AS AN INDICATOR OF VO_{2max}: A REPLICATION OF DISTANCE EFFECTS

R. R. Vickers, Jr.

20040204 193

Report No. 01-24

Approved for public release; distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P. O. BOX 85122
SAN DIEGO, CA 92186-5122

BUREAU OF MEDICINE AND SURGERY (MED-02)
2300 E ST. NW
WASHINGTON , DC 20372-5300

Running Performance as an Indicator of VO_{2max}: A Replication of Distance Effects

Ross R. Vickers, Jr.

Human Performance Program
Naval Health Research Center
P. O. Box 85122
San Diego, CA 92186-5122

e-mail: Vickers@nhrc.navy.mil

Report Number 01-24, supported by the Office of Naval Research, Arlington, VA, under Work Unit No. 63706N M0096.001-6417, and by the U.S. Army Medical Materiel and Research Command, under Army Reim-60109. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

This research has been conducted in compliance with all applicable Federal Regulations governing the protection of human subjects. No human subjects were directly involved in this review.

Approved for public release; distribution unlimited.

EXECUTIVE SUMMARY

Background

Running performance often is used to evaluate aerobic capacity. A previous review addressed the question "What is the relationship between distance and validity?" Bioenergetic models of running performance based on established physiological principles suggested that performance on longer runs would always yield more valid estimates than performance on shorter runs. The cumulative evidence from 122 studies contradicted this expectation. Validity increased with distance for shorter runs, but was constant for distances ≥2km. Also, validity was lower for fixed-time runs that were <12 min duration than for fixed-times ≥12 min.

Objective

This report was undertaken to determine whether the initial findings could be replicated.

Approach

The published literature was searched to identify studies of cardiorespiratory threshold measures (e.g., ventilatory threshold, anaerobic threshold) and running performance. A meta-analysis was conducted on reported correlations between VO_{2max} and performance extracted 74 correlations from 39 studies. The analyses cross-validated a set of statistical models initially developed and tested in the earlier review.

Results

The earlier findings replicated well. The model with increasing validity up to 2 km or 12 min and constant validity from those criterion points onward was the best representation of the data. An earlier finding that fixed-time run tests (e.g., a 12-min run) provided better estimates of aerobic capacity than fixed-distance run tests (e.g., 5 km) also replicated (r = .807 vs. r = .706).

Conclusions

Run tests should be at least 2 km in distance or 12 min in duration to maximize validity as indicators of aerobic capacity. Increasing distance or time beyond these minimum values does not improve run test validity as an indicator of VO_{2max}. Fixed-time tests have higher average validity than fixed-distance tests, so a 12-min run test will maximize validity while minimizing demands on the runners.

Running performance often is used to evaluate aerobic capacity. A previous review indicated that the validity of running performance increased with distance up to 2 km; validity was constant from 2 km onward (Vickers, 2001).[1]

Vickers's (2001) review was undertaken to address several questions. Do longer runs provide better estimates of aerobic capacity? Can the relationship between distance and validity be quantified? What is the shortest test that yields acceptable validity? How much is gained by increasing the test length beyond this minimum? The expectation was that the first two questions would be answered affirmatively and that the second two could be answered by constructing a simple mathematical model relating run distance to validity.

The anticipated answers to the questions posed in the initial review were based on empirical and theoretical considerations. Several studies have shown higher validity for longer runs (Burke, 1976; Farrell, Wilmore, Coyle, Billing, & Costill, 1979; Shaver, 1975; Weyand, Cureton, Conley, Sloniger, & Liu, 1994). Mathematical models of the bioenergetics of running provide a theoretical explanation for this trend (Capelli, 1999; di Prampero et al., 1993; Ward-Smith, 1999). These models suggest that validity will increase indefinitely with distance. However, the rate of increase will be slower as distance increases.

The review results were unexpected. The fact that validity only increased up to 2 km meant that there was a range of run distances for which the relationship was not strictly increasing as expected. Instead, the relationship would be characterized mathematically as nondecreasing. This observation is critically important when modeling the validity of run tests. No model, whether linear, curvilinear, or nonlinear, that predicts higher validity coefficients for longer tests will fit the data.

A piecewise (PW) model was formulated to represent the data. The model was piecewise because separate equations predicted validity for runs above and below the 2-km threshold. For runs less than 2 km, the predicted validity was determined by the logarithm of the distance. For runs of 2 km or longer, the prediction was a constant. Each range of predictions was one piece of the model.

[1] Validity is the appropriateness of the interpretation of a test score (American Psychological Association, 1985). Most tests can be interpreted more than one way and, therefore, have more than one validity. As used in this paper, validity refers solely to the interpretation of run test performance as an indicator of aerobic capacity. In this context, the term "validity coefficient" refers to the correlation between run test performance and maximal oxygen uptake capacity.

The PW model answered the original questions. If the minimum acceptable validity for a run test is *r* = .70, 2 km is the minimum run distance. The shortest fixed-time test would be 12 min. Adding distance or time to these minimum values does not increase test validity.

The prior review has two major implications. First, the results defined empirical criteria for classifying tests as endurance runs. The minimum criteria were 2 km or 12 min. Any run meeting either criterion is an endurance test. Although the validity of fixed-distance tests and fixed-time tests differed, the data indicated validity was constant within each category.

Second, the evidence provided an empirical basis for recommending one run test as the best option for estimating aerobic capacity. Fixed-time endurance tests produced higher validity coefficients than fixed-distance endurance tests. The reason for the difference was not clear, but fixed-time tests might increase the likelihood that runners will adopt the strategy of running at a constant pace throughout the test. This strategy yields optimal performance (Fukuba & Whipp, 1999). Whatever the basis for the difference, the best test for estimating aerobic capacity would be the shortest fixed-time endurance test. This test will yield the highest validity with the least effort and time required on the part of the test takers. The test also is valid for individuals who might have trouble running longer times or meeting a minimum distance requirement (Sidney & Shepard, 1977). Considering these criteria, the best run test would be a 12-min timed run.

Neither of the most important implications of the prior findings was anticipated when the review was undertaken. Unexpected findings should be viewed with skepticism until tested further. Replication is a constructive response to skepticism. This review, therefore, attempted to replicate the earlier work. The initial data set was extended by conducting a new literature search focused on physiological threshold variables as predictors of running performance rather than maximal aerobic capacity. The extended search identified 39 studies that reported 74 maximal oxygen uptake (VO_{2max}) running performance correlations that were not included in the initial review. These data were used to replicate and cross-validate the original findings.

Methods

Literature Search

The literature search had three primary elements. First, the PubMed® database was searched using "threshold" and "running" as the key words. The general term threshold was used in the hope that it would identify articles that dealt with various thresholds in the physiology literature. These included ventilatory threshold, lactate threshold, and anaerobic threshold.

Articles identified in the PubMed search were examined to determine whether they included useful data. The reference lists for articles that contained at least one useful correlation were examined to identify other studies that might report VO_{2max}-running performance correlations. The references identified in this process were compared with the list of articles covered in the earlier review (Vickers, 2001). New articles were examined to see whether they contained results that could be used in this review. This step comprised the ancestry review for the present work.

The reference catalog at San Diego State University was searched to identify dissertations and theses involving running. The list was compared with the citations in Vickers (2001) to determine which work had been examined previously. Those dissertations and theses not covered in the earlier review were examined to see whether they reported either correlations that would be used in this review or individual data that could be used to compute correlations.

The literature search identified 39 studies listed in Appendix A. These studies reported results from 50 distinct samples. The samples included 1,131 total participants who produced 1,769 running performance results. The outcome was a set of 74 correlations, 56 from published sources, including books. The other 18 correlations were from theses and dissertations. The average sample size for the 74 correlations was n = 23.9.

Data Extraction

The information extracted from each report consisted of the sample size, the type of run test (fixed-distance or fixed-time), the distance run, the average run time, and the VO_{2max}-running performance correlation. Performance was recorded a number of different ways in different studies. Performance on fixed-distance tests was usually recorded as a run time, but sometimes was represented by average running velocity. Performance on fixed-time tests typically was recorded as distance, but sometimes was reported as a predicted VO_{2max}. VO_{2max} predictions usually were computed using equations that involved only run distance. However, in some cases the predictions were based on multivariate equations with other predictors, such as weight or gender.

The signs of correlations with run time as the performance criterion were reversed so that correlations would have comparable meaning for all studies. For every other criterion, higher values indicated better performance. The correlations, therefore, were nearly all positive. In contrast, lower scores indicated better performance, and nearly all correlations were negative when run time was the criterion. Reversing the signs for these correlations meant that a positive correlation indicated how strongly VO_{2max} was related to good performance for all studies.

A separate record was constructed for each run test in a study. Thus, a study that included 1,500-m, 5-km, and 10-km runs produced 3 records, one for each distance. Sample attributes were duplicated on each record. Each record was treated as a separate case in the analysis. This decision meant that the cases analyzed were not entirely independent, thereby introducing statistical complexities for significance testing (Becker & Schram, 1994; Steiger, 1980). The common meta-analytic practice of averaging effect sizes to produce a single value for each sample was not suitable for the present purposes. Averaging would have prevented meaningful analysis of the relationship between validity and test length.

Analysis Procedures

As Rosenthal and DiMatteo (2001) noted, the underlying logic and basic computational procedures used in meta-analysis are the same as those used in the analyses of primary data. The basic summary statistics are weighted average correlations and computations of variance about those averages. In every analysis, the observed correlations are compared with predicted values based on the model. The estimated variance for the model provides a χ^2 test of statistical significance.[2]

The basic analysis followed the procedures in Chapter 11 of Hedges and Olkin (1985). Olkin and Pratt's (1958) formula was used to correct the correlations for sample size bias. Fisher's r-to-z transformation was applied to normalize the distribution of the corrected correlations (Hays, 1963). The transformed values are labeled $z_{UF(i)}$ to indicate that they represent *z* value of the <u>u</u>nbiased <u>F</u>isher-transformed correlation for the *i*th sample. The $z_{UF(i)}$ were the dependent variables in analysis of variance and regression procedures that weighted each observation by $(n_i - 3)$, where n_i is the sample size for the *i*th correlation. Using this weighting, the sums of squares reported for the analyses are χ^2 values that can be used to test hypotheses.

Three models were evaluated in both the replication and the cross-validation:

[2] Significance tests based on the χ^2 values should be interpreted with some caution given that not all of the validity coefficients were independent. However, only a small proportion of the total observations involved dependent coefficients. Note also that significance tests were not the basis for choosing the final model from the analyses.

A. The regression model used the logarithm of distance to predict $z_{UF(I)}$ values. This model is referred to as the LogDist model to indicate the predictor that was used in the regression.
B. The test-by-test (TxT) model predicted the average value for each of 9 groups. Seven groups represented specific distances represented in the data set by 3 or more correlations (1 mile, 2 km, 1.5 mile, 3 mile, 5 km, 10 km, marathon). Miscellaneous short (<1,850 m; *n* = 7) and long (>1,850 m; *n* = 9) runs were general groups that included all correlations for distances represented by just 1 or 2 correlations in the data set. This model was constructed using the same rules as the TxT model in Vickers (2001). The specific groups included differ because of differences in the data available for the analyses (see Appendix B for original model).
C. The PW model developed by Vickers (2001) regressed $z_{UF(i)}$ on the logarithm of distance for runs <2 km, then estimated a constant value for runs ≥2 km.

This review considered only these 3 models because they were the most promising of a larger set of models evaluated in Vickers (2001). The TxT model provided the best overall fit to the data in the initial review. This model minimizes the squared error in predictions for each run distance represented by 3 or more correlations. The TxT model, therefore, provides explanatory power that approaches the maximum possible value when distance is used as a predictor of validity. The LogDist model did not fit the data as well as either the PW or TxT models. However, the predicted values in this model increase continuously with distance. The rate of increase per unit distance decreases as distance increases. These attributes are characteristic predictions from bioenergetic models. Thus, this model was included as an approximation to predictions from bioenergetic models of running performance.

Vickers (2001) adopted the PW model over the TxT and LogDist models and several other models after weighing three criteria: explanatory power, number of parameters in the model (i.e., parsimony, cf., Popper, 1959), and relationships to physiological constructs. Considering the 3 models evaluated here, the regression model was simple and clearly linked to existing constructs but had the least explanatory power. The TxT model had the most explanatory power, but this model required many more parameters than either alternative model. Further, the pattern of mean differences as a function of distance was irregular and did not have a clear relation to physiological processes. The PW model provided intermediate explanatory power, but it combined parametric parsimony with a reasonable explanation in terms of known physiological mechanisms. Constant validity for endurance tests could be explained by concepts such as anaerobic threshold or critical power. The PW model also had the pragmatic value that it corresponded well to a simple graphic representation of the data.

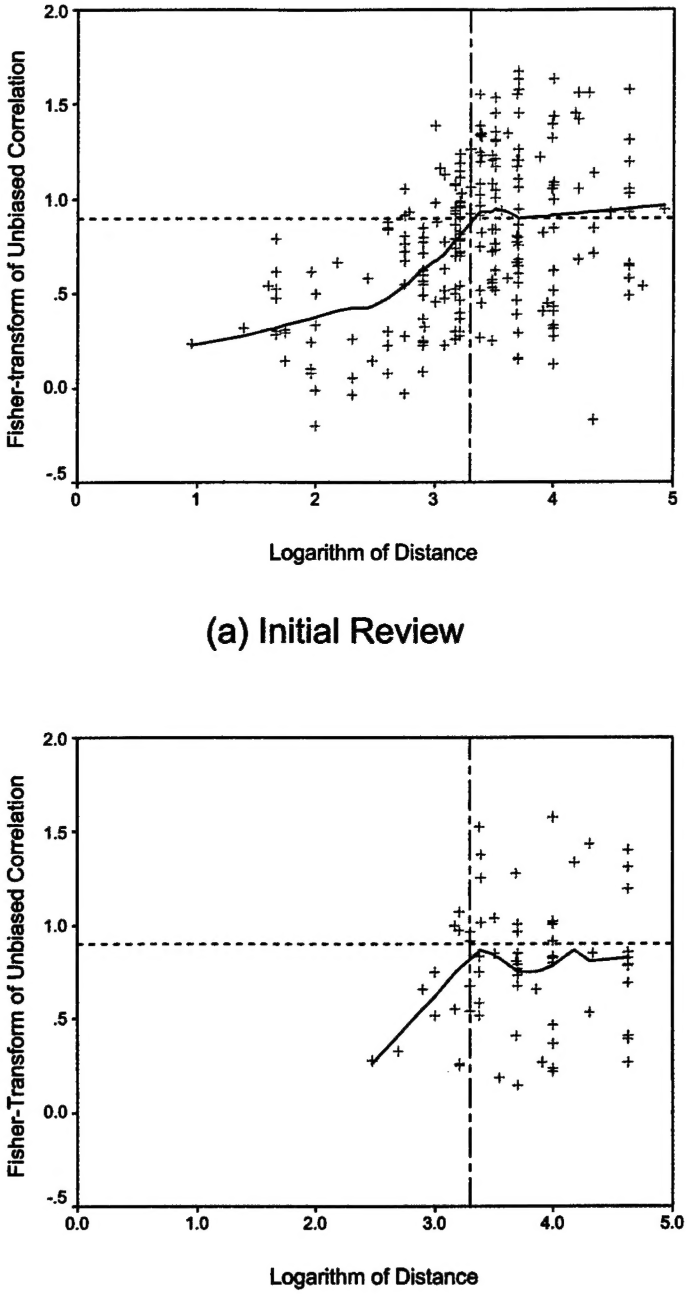
The 3 models retained from the initial review were compared in analyses that first replicated the original model selection process. For these analyses, parameter values for the models were estimated from the present data. The fit of each model was evaluated using the same criteria as in the initial review. This replication was undertaken to explore the possibility that the relative ordering of the models was specific to the initial data set.

The replication analysis was followed by cross-validation analyses. In these analyses, the model parameters were fixed at the values estimated in Vickers (2001). The parameter values are shown in Appendix B. The cross-validation represented an important shift in the work from exploratory analysis to confirmatory analysis.

All analyses were conducted using SPSS-PC (SPSS, Inc., 1998a,b). The weighted GLM and REGRESSION procedures were used. When correlations are appropriately transformed and weighted as previously described, the results include sums of squares that provide appropriate χ^2 values for testing meta-analytic hypotheses (Hedges & Olkin, 1985).

Parsimony-adjusted goodness-of-fit was used to compare models. The Tucker and Lewis (1973) index (TLI) was the basic goodness-of-fit indicator (cf., Arbuckle & Wothke, 1999; Bentler & Bonett, 1980; or Bollen, 1989; for discussion of goodness-of-fit indices). The TLI indicates what proportion of the greater than chance variation in correlations is accounted for by a model. Mulaik et al.'s (1989) parsimony adjustment was applied to the TLI to allow for the fact that more complex models almost always provide a better absolute fit to the data than simpler models. The final model criterion, therefore, was the parsimony-adjusted Tucker-Lewis Index (PTLI).

Hoelter's (1983) critical N was used to guard against assigning undue importance to small effects. Even trivial effects can be statistically significant given a large enough sample size (Rosenthal & Rosnow, 1984). Hoelter (1983) proposed that the potential for misleading significance tests be reduced by determining the smallest sample size for which an observed difference would be statistically significant. Hoelter (1983) labeled this sample size the critical N and suggested that any effect with critical N ≥ 200 was too small to be theoretically or practically important. This frame of references has been used when evaluating the present findings.



(a) Initial Review

(b) Current Data

Figure 1
LOESS Plots of Validity Coefficients as a Function of Distance

Table 1. Fit of Models Estimated From the Current Data

Model	df	Model χ^2	Residual χ^2	TLI	PTLI
LogDist	1	2.30	116.99	.010	.010
TxT	8	22.06	97.24	.163	.144
PW	2	11.76	107.53	.162	.157

Note. See text for model definitions. df = degrees of freedom for the model. Each model is based on 67 correlations (66 df maximum) with a total χ^2 of 119.30.

Results

Figure 1 plots $z_{UF(i)}$ as a function of the logarithm of distance for Vickers's (2001) data [Figure 1(a)] and the present data [Figure 1(b)]. The figure includes LOESS plots (Cleveland, 1979) for the data. The most important aspect of Figure 1 is that both LOESS plots are flat for distances ≥2 km. The horizontal reference line is Vickers's (2001) PW prediction for endurance runs (i.e., $z_{UF}' = .9026$). The flat portion of the LOESS curve for for the present data is slightly lower than, but approximately parallel to this reference line.

Figure 1 also shows increasing $z_{UF(i)}$ for runs <2 km in both data sets. This trend is poorly defined for the present data. Only a few data points are available for short runs. The available data points are largely restricted to the range of 1,500 m to 1,609 m. The two curves are similar for the data that are present.

Model Replication

Fitting the models to the data replicated the earlier findings (Table 1). The LogDist model again had the least predictive power. The variation explained by the PW model was statistically significant ($\chi^2 = 11.76$, 2 df, $p < .003$) but the TxT model explained more ($\chi^2 = 22.06$, 8 df, $p < .005$).

Some findings from the earlier review did not replicate. The LogDist model was statistically significant in the prior work, but not in these data ($\chi^2 = 2.30$, 1 df, $p > .129$). The TxT model was significantly better than the PW model in the prior review, but not in these analyses ($\Delta\chi^2 = 10.30$, 6 df, $p > .112$). Finally, the PW PTLI previously had been smaller than the TxT PTLI (.363 vs. .382), but was larger (.157 vs. .144) in these analyses.

Table 2. Cross-Validation Statistics for Model

Model	Residual χ^2	Difference χ^2	PTLI
LogDist	125.38	8.38	<.000
TxT	122.38	25.15	<.000
PW	109.70	2.168	.184

Note. All models have 67 df because no parameters were estimated. See text for definition of table entries.

Individual elements of the PW model replicated well. Shorter runs (i.e., <2 km) had lower average validity than longer runs (χ^2 = 9.09, 1 df, p < .001). The lack of association between distance and validity for longer runs (r = -.041, χ^2 = 0.172, 1 df, p > .678) replicated a second PW model element.

The only PW model component that did not replicate clearly was the significant relationship between distance and validity for short runs (χ^2 = 2.68, 1 df, p > .101, for the present data). However, the same trend was present and approached significance (p < .051) using a one-tailed test to allow for the fact that the direction of the relationship was known. Note also that there were only a few short tests (n = 13), most of which represented a narrow range of distance (9 of 13 either 1,500 m or 1,609 m). Taken in context, this element of the PW model replicated reasonably well within the constraints of the data.

Cross-Validation

Table 2 summarizes the results of the cross-validation analyses. The residual χ^2 values reported in the table are the result of fitting the corresponding Vickers (2001) model to the present data. The difference χ^2 is the difference between the cross-validation fit and the replication fit of the model (see Table 1). PTLI was computed for each model with the null model χ^2 = 119.30. This figure was the overall χ^2 for the set of validity coefficients. Model χ^2 values can be greater than this reference point because applying the parameter estimates from the earlier review to the present data can produce differences between predicted and observed correlations that are larger than the differences between the observed values and the mean correlation for the present data. The PTLI is negative when this outcome is obtained.

Cross-validation analyses clearly supported the PW model. First, the fit of the PW model (χ^2 = 109.70) was 12.5% better than the LogDist model (χ^2 = 125.38) and 10.4% better than the TxT model (χ^2 = 122.38).

Robustness of the model parameters was another indication of how well the PW model cross-validated. The difference in fit between replication and cross-validation was not significant for the PW model (χ^2 = 2.17, 3 df, p > .538). This result indicates that the original parameter values for the model were very close to the estimated values in the replication analysis. Sample-specific parameters were significantly better than the replicated values for the LogDist (χ^2 = 8.38, 3 df, p < .039) and TxT (χ^2 = 28.68, 14 df, p < .012) models.[3]

The goodness-of-fit index was the third reason to prefer the PW model to the alternatives. The PW PTLI was positive; the LogDist and TxT PTLI were negative. The negative PTLI indicated that bias in the cross-validation estimates (i.e., the tendency for predictions to err consistently in the same direction) was sufficient to offset whatever predictive power the models had for the new data.

The goodness-of-fit statistics produced another indication that the PW model was robust. The cross-validation PTLI was larger than the replication PTLI (.184 vs. .157). The reversal occurred because the lost degrees of freedom associated with estimating sample-specific parameters in the replication model more than offset the statistically nonsignificant gain in predictive accuracy.[4]

Detailed Cross-Validation of PW Model

The cross-validated PW model fit all of the data reasonably well. The most important element of the model was the prediction that z_{UF}' = .9026 for runs ≥2 km because these runs comprised

[3] The TxT cross-validation was based on predictions for 11 of 24 groups in the earlier review. The estimates from the earlier model were applied to all tests that fell in 1 of the 24 distance categories in that model. Thus, some tests classified as miscellaneous in the present review had distance-specific predictions.

[4] Negative PTLI values were obtained when cross-validation χ^2 >baseline χ^2. Biased cross-validation predictions produced this outcome. Bias is a consistent tendency toward underestimation or overestimation. The average bias (weighted by n - 3) was +.059 for the TxT model, +.037 for the LogDist model, and +.027 for the PW model. The biases were small [critical N (p < .05) 1,107, 2,810, and 5,273 for the TxT, LogDist, and PW models, respectively]and differed trivially. The critical N for the largest difference was 7,506. Bias added 10.70 to the PW χ^2, 11.90 to the Txt χ^2, and 16.35 to the LogDist χ^2.

5/6ths of the data. The weighted average in the present data was z_{UF}' = .8781. The critical N for the difference is 6,403. After back-transforming the *z* values, the difference in the estimated validity coefficients was *r* = .7176 versus *r* = .7055.

Predictions for short runs were less important for the overall fit of the model because there were fewer short runs. Here, too, the predictions were accurate. The 800-m prediction (n = 1 correlation) was slightly high (+.011). The 1-km (*n* = 2) prediction was slightly low (-.017). The 1-mile (*n* = 4) prediction was very close to the observed value (+.002). The largest discrepancy between observed and predicted value was +.062 for the 1,500-m run (*n* = 2). The critical N for the 1,500-m difference was 1,003. The critical N would be substantially higher for each other distance because of the smaller discrepancies. The overall model fit, therefore, reflected good fit at each cross-validated point.

Fixed Versus Random Effects

The replication analyses strengthened the choice of the PW model. Therefore, fixed- and random-effects versions of this model were compared. The fixed-effects model (χ^2 = 109.70) had slightly better predictive accuracy than the random-effects model (χ^2 = 115.18) when cross-validated.

Best Estimate Model

The prior analyses indicated that the fixed-effects PW model was the best representation of the data. The present data were combined with those from Vickers (2001) to estimate the parameters of that model using all of the data. The resulting PW model was:

If distance <2 km, z′ = (0.225*L) - .0615
If distance ≥2 km, z′ = .8960

where *z′* is the Fisher transformation of the unbiased correlation coefficient and L is the logarithm of distance. Pooling the data left the slope of the regression for short runs unchanged at 0.225. Pooling reduced the regression intercept slightly from the earlier value of -.0036 to -.0615. Pooling reduced the estimated value for long runs from .8960 to .9026 (critical N = 88,195). After back-transformation, the revised estimate of the validity coefficient was *r* = .714 compared with *r* = .718 in the initial review. The estimated validity coefficient applied equally well to all distances as indicated by the fact that distance and $z_{UF(i)}$ were independent (*r* = -.009) from 2 km upward.

Fixed-Time Tests

The data included 7 fixed-time correlations. Five of these correlations were for runs ≥12 min. The average correlation for those 5 tests was r = .807. This value was significantly (χ^2 = 6.42, 1 df, p < .012) larger than the average for fixed-distance tests in this review (r = .706). Both values were very similar to the estimates derived in the prior review (fixed-time, r = .793; fixed-distance, r = .718). Critical Ns for the differences between the two reviews were N > 2,542 for fixed-time tests and N > 6,508 for fixed-distance tests. The pooled significance for the difference between the fixed-time and fixed-distance correlations was p < .0001 by the method of adding probabilities (Rosenthal, 1978). The pooled average for fixed-time tests was r = .798.

Discussion

This extension of Vickers's (2001) review strongly supported the PW model relating run distance to validity. The LOESS plot of validity as a function of distance provides the most direct indication of support for the model. This graphic representation showed increasing validity for short runs and constant validity coefficients for long (i.e., ≥2 km) runs. These two trends are the essential elements of the PW model. Although this replication included relatively few short runs, the LOESS lines clearly were very similar.

Replication provided formal quantitative support for the original model selection process. The PW model once again had less predictive power than the TxT model and more predictive power than the regression model. Beyond this basic similarity, however, there were differences between the initial review and the present replication. Where Vickers (2001) found that the TxT model had significantly greater predictive accuracy than the PW model, the difference was not significant in this replication. Where Vickers (2001) found that the PTLI for the TxT model was slightly larger than the PTLI for the PW model, the replication reversed this ordering. Thus, 2 of 3 criteria that gave reason to consider choosing the TxT model over the PW model in the initial review were reversed in this replication. The only remaining criterion favoring the TxT model was the better absolute fit of the model to the data. The PTLI comparisons indicate that absolute fit is a weak criterion, given the substantial difference in complexity between the PW and TxT models. Applying the same criteria used in the earlier review, the results of this replication would lead to the adoption of a PW model.

Cross-validation underscored the replication trends. The explanatory power of the PW model was more than 10% greater than either competing model. The fact that the cross-validation fit of the PW model was nearly as good as the fit in the replication analyses indicated that the model parameters were robust. In fact, the difference in fit was not statistically significant, and the cross-validation PTLI for the PW model was larger than the simple replication PTLI for this model. The failure of the other two models to cross-validate is underscored by the fact that the PTLI was *negative* for both competing models.

These findings provide strong empirical support for the PW model. The cross-validation results are particularly noteworthy. These analyses provided a very strong test of the original models. The cross-validation test of a model was a stringent criterion because the parameters were fixed at specific values derived in the earlier review. This aspect of cross-validation analyses meant that a specific value was predicted for each validity coefficient in the cross-validation analyses. These point predictions increased the risk of failure for the model. Observed values had to be close to the specific predicted values to avoid a significant misfit between the data and the model. This requirement contrasts with null hypothesis testing procedures that treat any observed value that is significantly different from zero as support for a model. Cross-validation requires that the typical finding lie within the 95% confidence interval around the predicted value given the sample size. This confidence interval will be narrower than the range of all values that differ significantly from zero. The greater constraint on the range of data that indicate acceptable fit of the model makes the confirmatory cross-validation more likely to fail than the exploratory test of a null hypothesis model. In this sense, the cross-validation was a stronger test of the models (Meehl, 1990).

The risks associated with cross-validation were clearly evident in the results of these analyses. Two of the 3 models cross-validated so poorly that they had negative PTLI values. Only the PW model produced a positive PTLI. The fact that the fit of the cross-validated PW model was not significantly different from the fit of the PW model with sample-optimized parameter values further strengthened this model. This close fit between the data and specific point predictions for each observation in the data set is what Meehl (1990) refers to as "a darned strange coincidence." Such coincidence should strengthen faith in the model.

Several characteristics of the PW model could account for its cross-validation success. Parsimonious models provide more precise parameter estimates (Bentler & Mooijaart, 1989). Precise estimates should increase accuracy when applied to new data because they are less likely to be substantially different than the population parameters that are being estimated. A second point to consider is the fact that parsimonious models have less opportunity to capitalize on chance. Fewer parameters are fitted, so it is less likely that unnecessary parameters will be included by chance. The addition of parameters that represent chance trends in the initial data will increase error when applied to new data that do not include those chance trends. Having fewer parameters also reduces the likelihood that chance observations will lead to serious errors in the estimation of parameters that truly belong in the model.

One characteristic of the PW model that may have accounted for its cross-validation success is particularly noteworthy. The PW model was based on fitting mathematical functions to the data. This statement was true even for runs of 2 km or longer. In those cases, the function was a constant, but that constant was essentially the intercept in a regression analysis with a zero slope. As a result, the prediction for any given distance is influenced by the pattern of data for other run distances. This dependency on the overall pattern of data means the model "borrows strength" from other evidence in the data when estimating the value at a given spot (National Research Council, 1992). The borrowing effect should help correct errors that arise when just a few data points are available to estimate the correlation for a given run distance. In such cases, a single data point that was seriously in error could significantly bias the estimate for that distance. Fitting a function to the data yields estimates that smooth the curve by making the estimate consistent with nearby values rather than relying just on the data for that specific distance.

This review also replicated the difference between fixed-time and fixed-distance tests. Fixed-time endurance runs were more valid than fixed-distance endurance runs. The estimated validity for each type of test was very similar to that obtained in the prior review. On the whole, a fixed-time endurance test increases validity .084 relative to a fixed-distance endurance test (fixed-time, r = .798; fixed-distance, r = .714). The absolute difference is modest, but simple magnitude comparisons can be misleading. For example, if a run test were to be used to decide who meets a pass-fail criterion (e.g., 50th percentile of a distribution), the fixed-time test would classify 8.4% more people correctly (Rosenthal & Rubin, 1978).

The replication and cross-validation analyses reinforce the surprising answer to several questions addressed in Vickers's (2001) review. A 12-min run is the best option for estimating aerobic capacity. The standard error of estimate (SEE) for aerobic capacity using this test is ~3.8 ml/kg/min.[5] Laboratory VO_{2max} test precision is ~3.0 ml/kg/min when the same protocol is repeated twice or more (Froehlicher et al., 1974; Katch, Sady, & Freedson, 1982; Safrit, Hooper, Ehlert, Costa, & Patterson,

[5] $SEE = SD * \sqrt{(1 - r^2)}$ where SD = 6.24, the weighted average SD for all samples in the two reviews and r = .798, the average correlation between the 12-min run and the VO_{2max} assessments.

1988). Estimates from a 12-min run will have an SEE ~25% greater than the reference standard. The increase is not trivial, but it may be acceptable in many situations.[6]

This review suggests that one factor that might bias meta-analytic findings is unimportant in the present research domain. Increasing the scope of the literature review had little effect on the estimated validity coefficients. It is very unlikely that even the 347 correlations examined in the combined reviews exhaust the literature in this area. However, the fact that two different search strategies produced very similar results makes it less likely that omitted studies would change the results substantially. The reviews produced similar estimates despite differences in the proportional representation of published and unpublished studies. This outcome suggests that publication bias is not a major factor in this domain.

The preceding conclusion is subject to one critically important qualification. The inference is based on trends averaged across many types of people. The samples included males and females, young and old, and athletes and untrained individuals. Previous reviewers have cautioned that findings may not generalize across populations (Baumgartner & Jackson, 1982; Safrit et al., 1988). These cautions are still relevant. Figure 1 clearly shows that correlations vary widely for runs ≥2 km. Population differences may be one source of this variation.

The sources of variation in the validity of endurance run tests will be the topic of a companion review (Vickers, in preparation). This replication of Vickers's (2001) earlier findings sets the stage for a meaningful assessment of this topic by providing empirical criteria defining endurance runs. Runs ≥2 km or ≥12 min share a common validity within test type. The run test categories thus defined can both be classified as endurance runs. With this point established, analysis of the variation in validity coefficients for endurance runs can determine whether validity generalizes for endurance runs.

[6] The standard deviation specified for VO_{2max} tests applies to repetitions of a single protocol. Differences between protocols would be a more appropriate frame of reference. The reference SEE for that comparison would be larger because the SEE would include variance attributable to protocol differences.

References

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 user's guide.* Chicago: SmallWaters Corporation.

Baumgartner, T. A., & Jackson, A. S. (1982). *Measurement for evaluation in physical education.* (2nd ed.). Dubuque, IA: Wm. C. Brown.

Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357-382). New York: Russell Sage Foundation.

Bentler, P.M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Bentler, P. M., & Mooijaart, A. (1989). Choice of structural model via parsimony: A rationale based on precision. *Psychological Bulletin, 106*, 315-317.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Burke, E. J. (1976). Validity of selected laboratory and field tests of physical working capacity. *Research Quarterly, 47*, 95-103.

Capelli, C. (1999). Physiological determinants of best performances in human locomotion. *European Journal of Applied Physiology and Occupational Physiology, 80*(4), 298-307.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

di Prampero, P. E., Capelli, C., Pagliaro, P., Antonutto, G., Girardis, M., Samparo, P., & Soule, R. G. (1993). Energetics of best performances in middle-distance running. *Journal of Applied Physiology, 74*(5), 2318-2324.

Farrell, P. A., Wilmore, J. H., Coyle, E. F., Billing, J. E., & Costill, D. L. (1979). Plasma lactate accumulation and distance running performance. *Medicine and Science in Sports, 11*(4), 338-344.

Froelicher, V. F., Jr., Brammell, H., Davis, G., Noguera, I., Stewart, A., & Lancaster, M. C. (1974). A comparison of three maximal treadmill exercise protocols. *Journal of Applied Physiology, 36*, 720-725.

Fukuba, Y., & Whipp, B. J. (1999). A metabolic limit on the ability to make up for lost time in endurance events. *Journal of Applied Physiology, 87*, 853-861.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt, Rinehart, Winston.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. *Sociological Methods and Research, 11,* 325-344.

Katch, V. L., Sady, S. S., & Freedson, P. (1982). Biological variability in maximum aerobic power. *Medicine and Science in Sports and Exercise, 14,* 21-25.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108-141.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*(3), 430-445.

National Research Council. (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy Press.

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Popper, K. R. (1959). *The logic of scientific discovery.* NY: Basic Books.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193.

Rosenthal, R., & DiMatteo, R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. In S. T. Fiske, D. L. Schacter, & C. Zahn-Waxler (Eds.), *Annual review of psychology, Vol. 52* (pp. 59-82). Palo Alto, CA: Annual Reviews, Inc.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research.* New York: McGraw-Hill.

Rosenthal & Rubin (1978). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Safrit, M. J., Hooper, L. M., Ehlert, S. A., Costa, M. G., & Patterson, P. (1988). The validity generalization of distance run tests. *Canadian Journal of Sports Science, 13*(4), 188-196.

Shaver, L. G. (1975). Maximum aerobic power and anaerobic work capacity prediction from various running performances of untrained college men. *Journal of Sports Medicine, 15,* 147-150.

Sidney, K. H., & Shephard, R. J. (1977). Maximum and submaximum exercise tests in men and women in the seventh, eighth, and ninth decades of life. *Journal of Applied Physiology: Respiratory, Environmental, and Exercise Physiology, 43*(2), 280-287.SPSS, Inc. (1998a). *SPSS Advanced Statistics.* Chicago: SPSS, Inc.

SPSS, Inc. (1998b). *SPSS Base 8.0 Applications Guide.* Chicago: SPSS, Inc.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87,* 245-251.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

Vickers, R. R., Jr. (2001). *Running performance as an indicator of VO_{2max}: distance effects* (NHRC Tech. Rep. No. 01-20). San Diego, CA: Naval Health Research Center.

Ward-Smith, A. J. (1999). Aerobic and anaerobic energy conversion during high-intensity exercise. *Medicine and Science in Sports and Exercise, 31*, 1855-1860.

Weyand, P. G., Cureton, K. J., Conley, D. S., Sloniger, M. A., & Liu, Y. L. (1994). Peak oxygen deficit predicts sprint and middle-distance track performance. *Medicine and Science in Sports and Exercise, 26*(9), 1174-1180.

Appendix A

Data Sources

Bar-Or, O., Zwiren, D., & Dotan, R. (1978). Correlations among aerobic fitness tests and VO_{2max} in men who vary in aerobic power. In R. J. Shepard & H. LaVallee (Eds.), *Physical fitness assessment* (pp. 356-362). Springfield, IL: Charles C. Thomas.

Billat, V., Beillot, J., Jan, J., Rochcongar, P., & Carre, F. (1996). Gender effect on the relationship of time limit at 100% VO_{2max} with other bioenergetic characteristics. *Medicine and Science in Sports and Exercise, 28*(8), 1049-55.

Bulbulian, R., Wilcox, A. R., & Darabos, B. L. (1986). Anaerobic contribution to distance running performance of trained cross-country athletes. *Medicine and Science in Sports and Exercise, 18*(1), 107-113.

Chen, J. A. (1991). *Selected physiological variables and distance running performance among non-elite, heterogeneous groups of male and female runners.* Unpublished master's thesis, Washington State University, Pullman.

Claiborne, J. M. (1984). *Relationship of the anaerobic threshold and running performance in female recreational runners.* Unpublished doctoral dissertation, University of North Carolina at Greensboro.

Crain, M. L. (1977). *The relationship of brachial pulse wave components and predicted VO_{2max} to running performance.* Unpublished master's thesis, Northeast Missouri State University, Kirksville.

Di Prampero, P. E., Capelli, C., Pagliaro, P., Antonutto, G., Girardis, M., Zamparo, P., & Soule, R. G. (1993). Energetics of best performances in middle-distance running. *Journal of Applied Physiology, 74*(5), 2318-2324.

Dorociak, J. J. (1981). *Validity of running tests of 4, 8, and 12 minutes duration in estimating aerobic power for college women of different fitness levels.* Louisiana State University and Agricultural and Mechanical College, doctoral dissertation, Baton Rouge.

Foster, C. C., Jr. (1972). *Maximal aerobic power and the aerobic requirements of running in trained runners and trained non-runners.* Unpublished master's thesis, University of Texas at Austin.

Hagan, R. D., Upton, S. J., Duncan, J. J., & Gettman, L. R. (1987). Marathon performance in relation to maximal aerobic power and training indices in female distance runners. *British Journal of Sports Medicine, 21*(1), 3-7.

Haverty, M., Kenney, W. L., & Hodgson, J. L. (1988). Lactate and gas exchange responses to incremental and steady state running. *British Journal of Sports Medicine, 22*(2), 51-54.

Houmard, J. A., Costill, D. L., Mitchell, J. B., Park, S. H., & Chenier, T. C. (1991). The role of anaerobic ability in middle distance running performance. *European Journal of Applied Physiology and Occupational Physiology, 62*(1), 40-43.

Kitigawa, K., Yamamoto, K., & Miyashita, M. (1978). Maximal oxygen uptake, body composition and running performance in Japanese young adults of both sexes. In F. Landry & W. A. R. Orban (Eds.), *Exercise physiology: Fitness and performance capacity studies* (pp. 553-561). Miami, FL: Symposia Specialists, Inc.

Knowlton, R. G., & Gifford, P. B. (1972). An evaluation of a fixed time and fixed distance task as performance measures to estimate aerobic capacity. *Journal of Sports Medicine and Physical Fitness, 12*(3), 163-170.

Lambert, G. P. (1990). *The relationship between physiological measurements and cross-country running performance.* Unpublished master's thesis, Ball State University, Muncie, IN.

Laukkanen, R., Oja, P., Pasanen, M., & Vuori, I. (1992). Validity of a two kilmetre walking test for estimating maximal aerobic power in overweight adults. *International Journal of Obesity, 16,* 263-268.

Maughan, R. J., & Leiper, J. B. (1983). Aerobic capacity and fractional utilization of aerobic capacity in elite and non-elite male and female marathon runners. *European Journal of Applied Physiology, 52,* 80-87.

Mayers, N., & Gutin, B. (1979). Physiological characteristics of elite prepubertal cross-country runners. *Medicine and Science in Sports, 11*(2), 172-176.

Morgan, D. W., & Daniels, J. T. (1994). Relationship between VO_{2max} and the aerobic demand of running in elite distance runners. *International Journal of Sports Medicine, 15*(7), 426-429.

Morlang, C. (1988). *The effects of phosphate loading on red blood cell 2,3-DPG and endurance running performance.* Unpublished master's thesis, San Diego State University, CA.

Mosenthal, T. M. (1988). *Correlations of laboratory tests to distance running performance during a cross-country track season.* Unpublished master's thesis, St. Cloud State University, MN.

Myles, W. S., & Toft, R. J. (1982). A cycle ergometer test of maximal aerobic power. *European Journal of Applied Physiology and Occupational Physiology, 49*(1), 121-129.

Oja, P., Laukkanen, R., Pasanen, M., Tyry, T., & Vuori, I. (1991). A 2-km walking test for assessing the cardiorespiratory fitness of healthy adults. *International Journal of Sports Medicine, 12*(4), 356-362.

Porcari, J., Freedson, P., Ward, A., Rippe, J., Wilkie, S., Kline, G., Keller, B., & Hsieh, S. (1987). Prediction of VO_{2max}using the ACSM VO_2 prediction for running. *Medicine and Science in Sports and Exercise, 19,* S29.

Priest, J. W., & Hagan, R. D. (1987). The effects of maximum steady state pace training on running performance. *British Journal of Sports Medicine, 21*(1), 18-21.

Schrader, T. A. (1982). *Fluid ingestion and long-distance running.* Unpublished master's thesis, Arizona State University, Tempe.

Seljevold, P. J. (1989). *Prediction of running performance from selected variables measured during bicycle ergometry.* Unpublished master's thesis, St. Cloud State University, MN.

Takeshima, N., & Tanaka, K. (1995). Prediction of endurance running performance for middle-aged and older runners. *British Journal of Sports Medicine, 29,* 20-23.

Tanaka, K., & Matsuura, Y. (1984). Marathon performance, anaerobic threshold, and onset of blood lactate accumulation. *Journal of Applied Physiology, 57*(3), 640-643.

Tanaka, K., Matsuura, Y., Matsuzaka, A., Hirakoba, K., Kumagai, S., Sun, S. O., & Asano, K. (1984). A longitudinal assessment of anaerobic threshold and distance-running performance. *Medicine and Science in Sports and Exercise, 16*(3), 278-82.

Thiart, B. F., Blaauw, J. H., & van Rensburg, J. P. (1978). Endurance training and the VO_{2max} with special reference to validity of the Astrand-Ryhming nomogram and the Cooper 12-minute run as indirect tests for maximal oxygen uptake. In F. Landry & W. A. R. Orban (Eds.), *Exercise physiology: Fitness and performance capacity studies* (pp. 609-614). Miami, FL: Symposia Specialists, Inc.

Tokamakidis, S. P., & Leger, L. A. (1992). Comparison of mathematically determined blood lactate and heart rate "threshold" points and relationship with performance. *European Journal of Applied Physiology, 64,* 309-317.

Tokmakidis, S. P., Leger, L. A., & Pilianidis, T. C. (1998). Failure to obtain a unique threshold on the blood lactate concentration curve during exercise. *European Journal of Applied Physiology and Occupational Physiology, 77*(4), 333-342.

Walters, S. C. (1983). *The physiological effects of a twenty week distance running program on teenage girls correlated with performance.* Unpublished master's thesis, Arizona State University, Tempe.

Ward, A., Wilkie, S., O'Hanley, S., Trask, C., Kallmes, D., Kleinerman, J., Crawford, B., Freedson, P., & Rippe, J. (1987). Estimation of VO_{2max} in overweight females [Abstract]. *Medicine and Science in Sports and Exercise, 19,* S29.

Weltman, J., Seip, R., Levine, S., Snead, D., Rogol, A., & Weltman, A. (1989). Prediction of lactate threshold and fixed blood lactate concentrations from 3200-m time trial running performance in untrained females. *International Journal of Sports Medicine, 10*(3), 207-211.

Williams, K. R., & Cavanagh, P. R. (1987). Relationship between distance running mechanics, running economy, and performance. *Journal of Applied Physiology, 63*(3), 1236-1245.

Wiswell, R. A., Jaque, S. v., Marcell, T. J., Hawkins, S. A., Tarpenning, K. M., Constantino, N., & Hyslop, D. M. (2000). Maximal aerobic power, lactate threshold, and running performance in master athletes. *Medicine and Science in Sports and Exercise, 32*(6), 1165-1170.

Yoshida, T., & Ishiko, T. (1978). Physiological studies on cardiorespiratory response to exercise validity of endurance tests in ten-year-old boys. In F. Landry & W. A. R. Orban (Eds.), *Exercise physiology: Fitness and performance capacity studies* (pp. 541-545). Miami, FL: Symposia Specialists, Inc.

Appendix B

Cross-Validation Models

Logarithm of Distance (LogDist) Model:

$z_{UF(i)}' = 0.240*L -.013$

Test-by-Test (TxT) Model:

Distance (m)	$z_{uf(i)}'$
800	.500
1000	1.024
1500	.744
1609	.732
2000	1.017
2414	1.075
3200	.832
4827	.806
5000	.932
10000	.715
21100	.880
42200	1.015
Misc Short	.---
Misc Long	.---

Piecewise (PW) Model:

If (distance < 2000 m) $z_{UF(i)}' = 0.225*L - .036$
If (distance ≥ 2000 m) $z_{UF(i)}' = .903$

Note. "L" indicates the logarithm of distance in meters.

REPORT DOCUMENTATION PAGE

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB Control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. Report Date (DD MM YY)
23 Jul 01

2. Report Type
Interim

3. DATES COVERED (from - to)
March to July 20013

4. TITLE AND SUBTITLE
Running Performance as an Indicator of VO$_{2max}$: A Replication of Distance Effects

5a. Contract Number: US Army Reimbur-
5b. Grant Number: sable-60109
5c. Program Element: 63706N
5d. Project Number: M0096
5e. Task Number: 001
5f. Work Unit Number: 6417
5g. IRB Protocol Number:

6. AUTHORS
Ross R. Vickers, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

8. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

Office of Naval Research
800 North Quincy St
Arlington, VA 22217-5600

Bureau of Medicine and Surgery
M2
2300 E St NW
Washington DC 02372-5300

9. PERFORMING ORGANIZATION REPORT NUMBER
Report No 01-24

10. Sponsor/Monitor's Acronyms(s)
ONR/BUMED

11. Sponsor/Monitor's Report Number(s)

12 DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT (maximum 200 words)

An earlier review showed that run test validity as an indicator of aerobic capacity increased with test distance up to 2 km and with test duration up to 12 min, then remained constant. This representation of the relationship between distance/time and test validity was labeled the piecewise (PW) model of run test validity. The range of constant validity in the PW model was surprising in light of physiological and mathematical models of running performance. Therefore, this review analyzed 74 correlations from 39 additional studies to replicate the prior findings if possible. Cross-validation of the models showed that only the PW model had positive predictive value in the new data. An earlier finding that fixed-time run tests (e.g., a 12-min run) provided better estimates of aerobic capacity than fixed-distance run tests (e.g., 5 km) also replicated (r = .807 vs. r = .706). Based on this evidence, a 12-min run test maximizes validity while minimizing demands on the runners.

15. SUBJECT TERMS
run tests, aerobic capacity, validity, modeling

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Commanding Officer
a. REPORT	b.ABSTRACT	b. THIS PAGE			
UNCL	UNCL	UNCL	UNCL	25	19b. TELEPHONE NUMBER (INCLUDING AREA CODE) COMM/DSN: (619) 553-8429